DOCUMENT RESUME

ED 211 578 TM 810 961

AUTHOR Rodger, R. S.
TITLE A Classification and Comparison of Multiple Comparisons.
PUB DATE [74]
NOTE 11p.; Paper presented at the Annual Meeting of the Canadian Psychological Association (Toronto, Canada, June 3-5, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; *Evaluation Methods; *Hypothesis Testing; Mathematical Models; Research Methodology; *Statistical Analysis
IDENTIFIERS *Error Analysis (Statistics); *Multiple Comparisons

ABSTRACT

A fairly large number of different methods exist for evaluating comparisons or null hypothetical contrasts. Because it is difficult to choose among them, this paper develops a 2 by 2 classification of the methods. The first dimension of the classification is decision-based error rate vs. experimentwise error rate. The second dimension is planned vs. post-hoc contrast. Most of the generally useful and better-known methods are considered: t, Bonferroni, Dunnett, Duncan Multiple Range, Rodger, Scheffe, Tukey, and Newman-Keuls Multiple Range. Each method is classified and discussed. If one has very specific questions to be answered, and a very good idea of what the true population means are, use carefully planned tests; otherwise use post-hoc methods. If one believes that a null hypothesis of equal means may be true and that one should have a conventionally low probability of rejecting it in error, use experimentwise error rate, but keep the number of means being considered as small as possible; otherwise use decision-based error rates. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED211578

A Classification and Comparison of Multiple Comparisons

R. S. Rodger

Dalhousie University

A fairly large number of different methods exist for evaluating comparisons or null hypothetical contrasts and new techniques are still being added to the literature. Many of the methods were described by Miller (1966) and Miller (1977) gave an almost complete bibliography of research on the subject up to that date. But reviews and overviews of this field seem to add more confusion than clarification for the casual user. Faced with the plethora of procedures it is difficult to choose among them without a rationale to guide one. This paper develops a 2 by 2 classification of the methods in the hope that this will provide some basis for choice.

Most of the generally useful and better-known methods are considered, i.e., t, Bonferroni, Dunnett, Duncan Multiple Range, Rodger, Scheffé, Tukey and the Newman-Keuls Multiple Range methods. They are compared to one another with respect to how large $\Sigma c_j m_j$ must be to reject the hypothesis $\Sigma c_j \mu_j = 0$: the c_j are the contrast coefficients, the m_j are the sample means and the μ_j are the true population means. For all except the classical t-test the size required increases as the number of groups (or number of groups -1) increases, but the rate of increase differs from method to method.

The size of $\Sigma c_j m_j$ required to reject $\Sigma c_j \mu_j = 0$ cannot be considered in isolation from the other properties of the procedures. For example, it should be no surprise to learn that the critical size of $\Sigma c_j m_j$ is related to the freedom one has to choose values for c_j, and to whether these values must be set independently of how the m_j turn out (choosing the c_j "blindly" so to speak) or as a function of the observed sample means. Some grasp of these relationships may be obtained from the report of random sample data for J = 10 groups of N = 2 observations each, replicated throughout 100 monte-carlo experiments. These data are analysed by the various methods and the number of null hypothetical contrasts $\Sigma c_j \mu_j = 0$ each rejects is reported, as well as the number of experiments in which at least one such null is rejected.

Finally, the various procedures differ in their mathematical elegance, simplicity and generality. These features are discussed at the end of the paper.

TM 810 961

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. S. Rodger
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Planned versus Post-hoc Contrasts.

Multiple-comparison methods must be classified as either suitable only for planned tests (i.e., when the contrast coefficients c_j are chosen by the investigator independently of the sample means m_j and variance s^2 used to test these nulls, preferably before such data are collected), or as suitable for evaluating contrasts selected in the light of how the sample data m_j and s^2 have turned out (i.e., post-hoc procedures). Many writers tend to ignore this crucial distinction (e.g., Miller, 1966) and some (e.g., Ryan, 1959, p. 27-28) have argued that it is slight or non-existent (but see Gaito, 1959). It is, however, easy to show that the type I error-rate claimed for some methods may be very wide of the mark unless the null contrasts considered are planned (see Rodger, 1973 for how this applies to the use of Bonferroni's inequality).

Obviously, the reader of reports of planned-tests takes the honesty and competence of the planned-tests-reporter on faith. Such faith would be higher if planned-tests-reporters stated explicitly that they had planned the only contrasts to be tested before collecting their sample data, or if the readers had not heard so many research workers claim that a contrast which turns out to be so interesting post hoc could easily have been "planned" if they had just thought of it!

Decision-based versus Experimentwise Error-rate

The second basis for classification is the definition of type I error-rate. A variety of bases are possible: decision-based, comparisonwise, hypothesis-wise, per-degree-of-freedom, experimentwise and familywise are all terms which have been used here. In practice, the distinction is between the decision basis for error-rate or the experimentwise basis. The former, consistent with the Neyman and Pearson (1928) theory of testing statistical hypotheses, sets the type I error-rate as the probability of rejecting a true null hypothesis in making a decision about it, or the expected relative frequency with which true nulls are rejected among all true nulls decided. The experimentwise error-rate I is the probability of rejecting at least one true null contrast in an experiment in which all null contrasts are true, or the expected proportion of experiments, for which all null contrasts are true, in which at least one true null is rejected. When H_0: $\mu_1 = \mu_2 = \ldots = \mu_j$ is true, then all null contrasts $\Sigma c_j \mu_j = 0$ are true (c_j not all zero and $\Sigma c_j = 0$). Of course, if one sets the type I error-rate for rejecting the above H_0 in error, and if that is the only decision to be made, then the experimentwise error-rate is the

decision-based error-rate for H_0.

At this point it is important to note that if the decision-based error-rate is set at αd it is often possible to compute the experimentwise rate (αe) arising therefrom, and vice versa. To illustrate this, a formula which sometimes applies, at least as a crude bound, is

$$\alpha e = 1 - (1 - \alpha d)^k; \qquad (1)$$

so for $\alpha d = .05$ with $k = 4$ decisions $\alpha e = .185$, and for $k = 8$ $\alpha e = .337$: vice versa for $\alpha e = .05$ with $k = 4$ then $\alpha d = .013$, and $k = 8$ gives $\alpha d = .006$. When a reference is made in this paper, therefore, to setting error-rate I on a decision basis or experimentwise it must be understood that the rate being set is a conventional one (i.e., .05 or .01).

Another matter which affects the interpretation of error-rate is whether it refers to decisions made by the investigator in the privacy of his or her office or whether it refers to results reported in published papers and/or at meetings of learned societies. This last definition might better be called "the publication error-rate" to stand beside "the presentaion error-rate" and "the casual conversation error-rate". The first (personal) definition will be used here: that we can control, the others being influenced by editorial boards, editors, referees and numerous social facilitators and inhibitors!

Finally, how many decisions shall be made, how will they be chosen and where does an experiment end and another begin? To cut through much possible debate, let there be decisions made for the maximum number of linearly independent contrasts (for example, that is J-1 for J groups) and let the decisions be to reject or accept (or not reject) the null forms in accordance with the rejection rule (which should be stated before the data are collected).

An experiment could be defined as a set of contrasts (or the space spanned by them) for which some scientific theory or model implies that all nulls are true. This could be an excellent definition of a family of statistical hypotheses and setting the experimentwise error-rate would then be equivalent to setting the familywise error-rate. If even one null contrast was false, then the scientific theory would be false and one might make a case for using the theorywise (or familywise) error-rate. In fact, in so far as experiments test theories at all, they do not tend to test them in this all or none fashion. On the rare occasions when this is the intention, the theoretically expected pattern of results is usually not all zeroes and often

the expected pattern spreads through an analysis of variance or two, a few contingency chi-squared tests and the occasional few correlations. Nobody has seriously suggested subjecting such a conglomerate to a single statistical test procedure. The classical origin of experimentwise error-rate may be the notion that null hypotheses are hypotheses of "no effect", but that interpretation of nulls is not universally accepted and the notion itself hardly qualifies as a scientific theory. In practice, an experiment is taken to be a set of data which yield a single test statistic: such as F in anova; or each of F_I, F_J and F_{IJ} in a factorial analysis; or a chi-squared value; or a manova test statistic. This gives one great freedom. Thus in a complicated multi-factorial design one may be able to combine all results into a single overall F-ratio, or partition the results into a series of main-effect and interaction F-ratios or, in the limit, into a series of planned contrasts with an F-ratio for each. In this last case an experiment is a contrast and the experimentwise error-rate becomes synonymous with the decision-based error-rate. The concept of experimentwise error rate might be better reformulated to refer to the entire pattern of results one wishes to claim is true (including both accepted and rejected nulls, as well as sizes asserted for the latter) as opposed to an alternative and different outcome. In this context it is important to note that when the true μ_j differ from one another there are still numerous (∞ when $J > 3$) null contrasts across these μ_j which are true. Little work has been done on this and the traditional definition of experimentwise error-rate will be used here.

Classification and Comparison

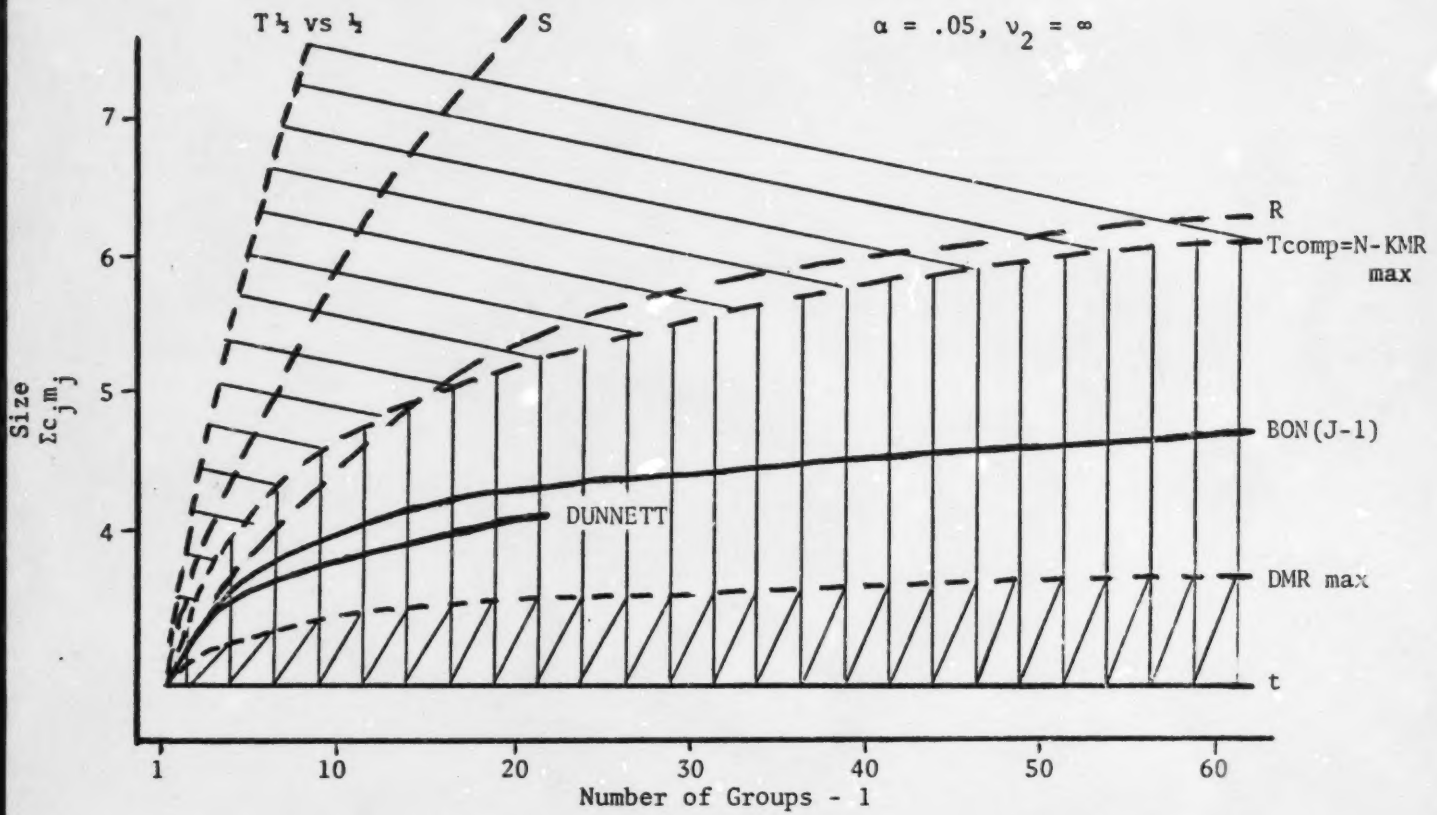
The eight techniques considered are classified in Table 1.

Table 1

	Planned	Post-hoc
Decision-Based	$t = \sqrt{F_{\cdot 05;1,\infty} * 2}$	$DMR = q_{\cdot 95;i,\infty}^{i}$ $R = \sqrt{(J-1)F_{[\cdot 05];J-1,\infty} * 2}$
Experimentwise	$BON = \sqrt{F_{\cdot 05/(J-1);1,\infty} * 2}$ $DUN = mt_{\boxed{\cdot 95};1,\infty}\sqrt{2}$	$N\text{-}KMR = q_{\cdot 95;i,\infty}$ $T = (q_{\cdot 95;J,\infty})\frac{1}{2}\Sigma\lvert c_j\rvert$ $S = \sqrt{(J-1)F_{\cdot 05;J-1,\infty} * 2}$

For purposes of comparability, the value $\Sigma c_j^2 = 2$ has been used throughout Table 1 and for the graph lines in Figure 1.

Figure 1



It should be noted that DMR, N-KMR and Tukey techniques each cover a range on Figure 1. The DMR line shown represents the amount required to reject $\mu_{\text{largest m}} - \mu_{\text{smallest m}} = 0$ and, if that is rejected, DMR becomes more generous. If rejections continue, DMR ends up looking at $\mu_{(j+1)} - \mu_{(j)} = 0$, i.e., at adjacent ordered m_j, and there DMR = t. The same is true for N-KMR. As is well-known, Tukey's method is most generous for comparisons (the lower T line) and least generous for comparing half of the groups of J means with the other half (the upper T line). Other types of contrasts fall between these two limits with Tukey's method.

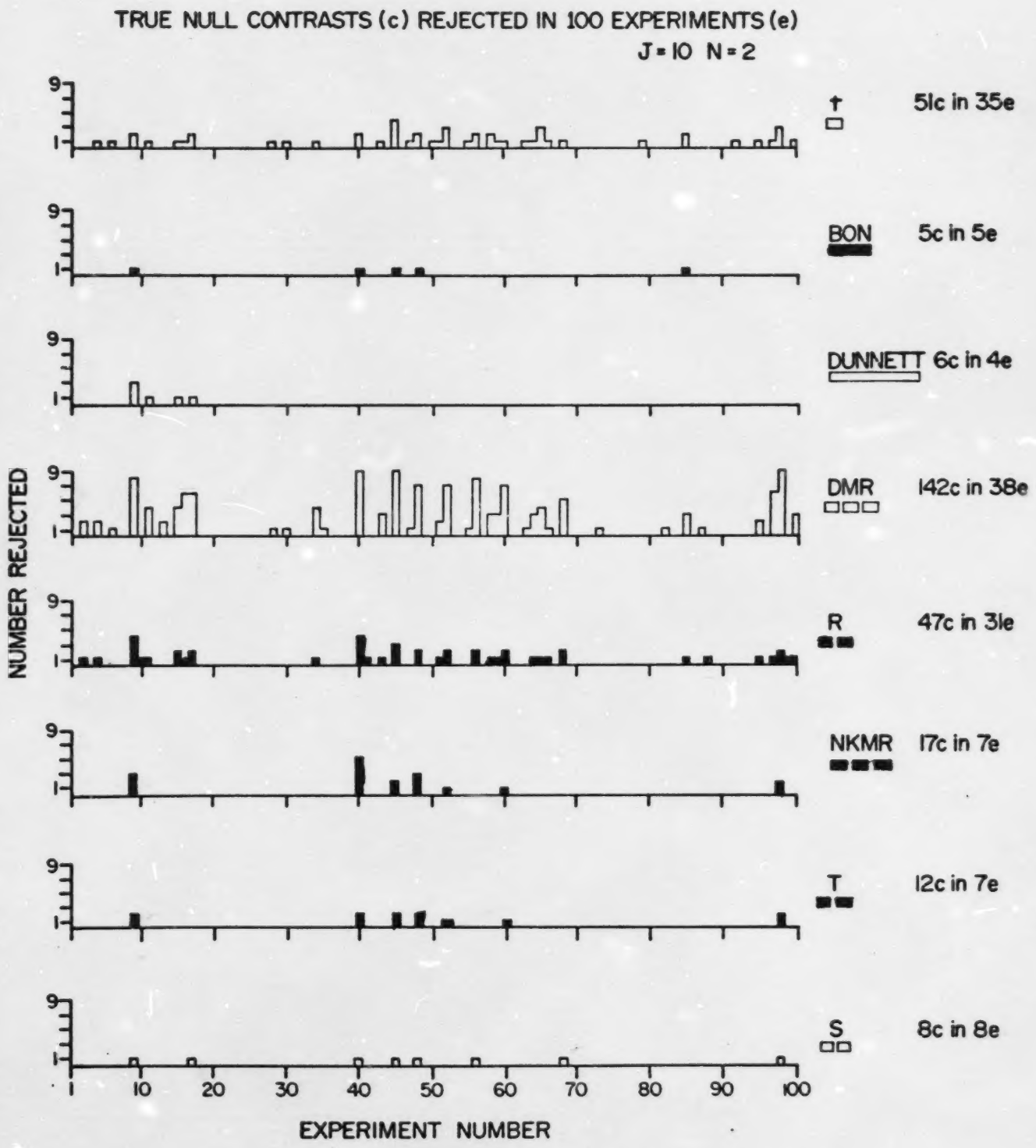
Figure 1 might lead one to suppose that the t-test is the thing to use because it is most generous. Actually it is no more generous than its competitors with decision-based error-rates, i.e., DMR and Rodger. This is because of the freedom these last two methods have to select comparisons or

contrasts for decision making in the light of how the sample means m_j and variance s^2 have turned out. This is shown in the monte-carlo study reported in Figure 2. There were 100 experiments carried out, each of J = 10 groups and each of those with N = 2 observations drawn at random. The number of null comparisons (for DMR, N-KMR and Dunnett) or mutually orthogonal null contrasts (for t, Bonferroni, Rodger, Scheffé and Tukey) rejected is given. Since J - 1 = 9, the expected total number of nulls rejected in $900\alpha d = 45$. Both t(51) and R(47) are very close to that, DMR(142) is much too large. The DMR discrepancy arises from technique's control of the tail of the error-rate distribution (i.e., it sets $\alpha d = .05$ in (1) then uses the resulting $\alpha e = .370$) rather than its expectation and the fact that DMR is a step-down procedure becoming ever more generous as outer null comparisons are rejected.

With the experimentwise error-rate at .05, the expected number of experiments in which one or more null contrasts will be rejected is $100\alpha e = 5$. The monte-carlo results for Bonferroni (5), Dunnett (4), Scheffé (8), Tukey (7) and N-KMR (7) are very close to that.

Research workers naturally do not like to make errors; so maybe rather than t being the method of choice one might prefer the Bonferroni, Dunnett, Scheffé or Tukey methods since they reject very few true null contrasts in error. Unfortunately, closing the doors on errors of one kind (type I) simply opens the windows to let in errors of the other kind (type II) or requires substantial increases in sample size N as the number of groups J increases. Thus if $\Sigma(\mu_j - \mu_.)^2/\sigma^2 = 1$ and $\beta = .95$ is the desired power (i.e., the probability of correctly rejecting H_0: $\mu_1 = \mu_2 = \ldots = \mu_J$), then with $\alpha e = .05$ Scheffé's method requires N = 13, 24, 31, 40 and 47 for J - 1 = 1, 9, 19, 39 and 59 respectively. Since the Tukey and N-KMR procedures use the Studentized Range as their test statistic, their power β depends not only on the amount of variation in the μ_j, i.e., on $\Sigma(\mu_j - \mu_.)^2/\sigma^2$, but also on their pattern (see David, Lachenbruch and Brandis, 1972). For the most powerful pattern let $\mu_1 = -\sigma/\sqrt{2}$, $\mu_J = \sigma/\sqrt{2}$ and all other $\mu_j = 0$. In that case $\mu_. = 0$ and $\Sigma(\mu_j - \mu_.)^2/\sigma^2 = 1$ as before. Now for $\beta = .95$ the required N = 13, 19, 22, 29 and 31 for J = 2, 10, 20, 40 and 60 respectively. Of course, if $\Sigma(\mu_j - \mu_.)^2/\sigma^2 = 1$ and the μ_j were spread unequally between μ_1 and μ_J the required sample size N for $\beta = .95$ would be greater than those quoted. The worst case would be half of the μ_j having the value $-\sigma/\sqrt{J}$ and the other half the value $\sigma/\sqrt{J}$:

Figure 2


TRUE NULL CONTRASTS (c) REJECTED IN 100 EXPERIMENTS (e)
J = 10 N = 2
NUMBER REJECTED
EXPERIMENT NUMBER
1
10
20
30
40
50
60
70
80
90
100
9
t
51c in 35e
BON
5c in 5e
DUNNETT
6c in 4e
DMR
142c in 38e
R
47c in 31e
NKMR
17c in 7e
T
12c in 7e
S
8c in 8e

for example, this requires N = 33 for J = 10 to make $\beta = .95$ with $\alpha e = .05$.

Decision-based error-rates do not give rise to loss of power, or to greatly increased sample size N to maintain it, as the number of groups J increases. Rodger (1974, p. 195) showed that for a fixed non-central parameter, the power β actually increases somewhat as J increases if his method is used. Of course, decision-based error-rate users might prefer to set rejection-rate per decision rather than the experimentwise power β. For planned t, N = 13 would remain constant as J increases with $\alpha d = .05$ and $\Sigma c_j \mu_j = \sigma\sqrt{\Sigma c_j^2}$. With Rodger's method N = 13, 10, 13, 15 and 16 is required for $\alpha d = .05$ (or $[E\alpha] = [.05]$ as he defines it) as J - 1 = 1, 9, 19, 39 and 59 respectively when $\Sigma c_j \mu_j = \sigma\sqrt{\Sigma c_j^2}$.

Elegance, Simplicity and Generality.

There is no doubt that the variance-ratio methods (t, Bonferroni, Rodger and Scheffé) have greater generality than studentized-range methods. Thus the former apply in all fixed-effects analysis of variance contexts even with unequal sample sizes. An approximation can be used with studentized-range methods when the sample sizes are unequal, but then only for comparisons. Variance-ratio methods can be used to evaluate linear hypotheses which are not contrasts. Since tables of the augmented studentized range are not available, studentized-range methods for evaluating such hypotheses are only approximate. Variance-ratio methods (or their multivariate generalizations) can be used when the implicit variances are unequal (e.g., in analyses of covariance, Hotelling's regular T^2 and in manova). Studentized-range methods are not then appropriate. Variance-ratio methods have been adapted to problems in multiple regression and correlation; studentized-range methods do not apply here.

Variance-ratio methods use a single non-central parameter (Δ) when power is computed and that is defined in a manner analogous to F, i.e.,

$$F = N\Sigma(m_j - m_.)^2/\nu_1 s^2 \tag{2}$$

$$\Delta = N\Sigma(\mu_j - \mu_.)^2/\sigma^2 \tag{3}$$

$$F = N(\Sigma c_j m_j)^2/\nu_1 s^2 \Sigma c_j^2 \tag{4}$$

$$\Delta = N(\Sigma c_j \mu_j)^2/\sigma^2 \Sigma c_j^2 \tag{5}$$

$$F = N_1 \nu_H ({}_H C_{JJ} C_H^T)^{-1} {}_H \nu_1^T/\nu_1 s^2 \tag{6}$$

$$\Delta = N_1 \delta_H ({}_H C_{JJ} C_H^T)^{-1} {}_H \delta_1^T/\sigma^2 \tag{7}$$

Power for studentized-range methods depends on the pattern of the μ_j as well as the amount of their variation.

Dunnett's procedure, though it is a variance-ratio method, is restricted to one particular pattern of contrasts (i.e., the many-one set) and multiple-range methods are restricted to comparisons. The restriction to comparisons is more serious than it may seem; for example, if one value of m_j falls between two others (say m_i and m_k) it is not possible to evaluate $\mu_j - (\mu_i + \mu_k)/2 = 0$, which asserts that μ_j falls midway between μ_i and μ_k.

The planned tests methods are limited in their usefulness to those situations in which some specific pattern or patterns among the μ_j can be strongly anticipated. Without good prior knowledge of how the μ_j are likely to turn out (assuming not all μ_j are equal), the contrasts selected for testing may miss the most powerful patterns. For example, if $\mu_j = -3\sigma/2, -\sigma/2, \sigma/2, 3\sigma/2$ then a t-test with N = 6 and $\alpha d = .05$ gives

$$\begin{array}{lll} \mu_3 - \mu_1 = 2\sigma & \Delta = 2N = 12 & \beta \doteqdot .92 \\ \mu_4 - \mu_2 = 2\sigma & \Delta = 2N = 12 & \beta \doteqdot .92 \\ \mu_4 + \mu_2 - \mu_3 - \mu_1 = 2\sigma & \Delta = N = 6 & \beta \doteqdot .70 \end{array} \qquad (8)$$

A different set of planned contrasts gives

$$\begin{array}{lll} \mu_2 - \mu_1 = \sigma & \Delta = N/2 = 3 & \beta < .5 \\ \mu_3 - \mu_2 = \sigma & \Delta = N/2 = 3 & \beta < .5 \\ \mu_4 - \mu_3 = \sigma & \Delta = N/2 = 3 & \beta < .5 \end{array} \qquad (9)$$

Decisions for the last set are likely to miss the mark more often than those for the former. (For a witty comment on such matters one should read the poem "Hiawatha Designs an Experiment" by Kendall, 1959.)

Conclusion.

If one has very specific questions to be answered, and one has a very good idea of what the true μ_j are, use carefully planned tests, otherwise use post-hoc methods. If one believes that H_0: $\mu_1 = \mu_2 = \ldots = \mu_J$ may be true and that one should have a conventionally low probability of rejecting it in error, use experimentwise error-rate but keep J as small as possible, otherwise use decision-based error-rates.

References.

David, H. A., Lachenbruch, P. A. and Brandis, H. P. The power function of range and studentized range tests in normal samples. Biometrika, 1972, 59, 161-168.

Gaito, J. Multiple comparisons in analysis of variance. Psychol. Bull., 1959, 56, 392-393.

Kendall, M. G. Hiawatha designs an experiment. American Statistician, 1959, 13, 23-24.

Miller, R. G. Simultaneous Statistical Inference. McGraw-Hill: New York, 1966.

Miller, R. G. Developments in multiple comparisons 1966-1976. J.A.S.A., 1977, 72, 779-788.

Neyman, J. and Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika, 1928, 20A, 175-240 and 263-294.

Rodger, R. S. Confidence intervals for multiple comparisons and the misuse of the Bonferroni inequality. Brit. J. math. statist. Psychol., 1973, 26, 58-60.

Rodger, R. S. Multiple contrasts, factors, error rate and power. Brit. J. math. statist. Psychol., 1974, 27, 179-198.

Ryan, T. A. Multiple comparisons in psychological research. Psychol. Bull., 1959, 56, 26-47.